KI - Künstliche Intelligenz
https://doi.org/10.1007/s13218-017-0520-x

TECHNICAL CONTRIBUTION

Societal Implications of Big Data

Karolin Kappler[1] · Jan-Felix Schrape[2] · Lena Ulbricht[3] · Johannes Weyer[4]

Received: 28 April 2017 / Accepted: 4 December 2017

Abstract
Modern societies have developed a variety of technologies and techniques to identify, measure and influence people and objects. Smart devices such as smartphones and wearables assist and track their users in every aspect of life. Large amounts of data are collected, evaluated and interconnected to analyse the behaviour of individuals, social groups and collectives. By discussing recent practices of self-tracking as well of real-time control of complex systems, we will show that real-time analysis and feedback loops increasingly foster a society of (self-)control. Data scientists and social scientists should work together to develop the concepts of regulation, which are needed to cope with the challenges and risks of big data.

1 Introduction

Modern societies have developed a variety of technologies and techniques to identify, measure and influence people and objects. Smart devices such as smartphones and wearables assist and track their users in every aspect of life. Large amounts of data are collected, evaluated and interconnected to analyze the behavior of individuals, social groups and collectives.

Since 2015, a group of German social scientists has conducted the research project "Assessing Big Data" (ABIDA)[1] with the objective to analyze the societal implications of Big Data—particularly elaborating the question, in what way and to what extent modern technologies of data analysis entail novel societal benefits and risks that differ from previous socio-technical configurations. This paper sets out to discuss a couple of theses concerning this question.

After a short retrospective of digital data analysis, we lay out various applications of data analytics in the light of distinct research cultures between data science and social science. Subsequently, we elaborate how self-tracking and real-time control of complex systems on the basis of data analytics create new individual and collective practices that shape modern societies. As these practices bear many risks for individual rights and collective achievements, we finally shed light on various propositions for the regulation of Big Data. Our argument is that Big Data, real-time analysis and feedback loops increasingly foster a society of (self-)control. But at the same time, these technologies are highly contentious and set the stage for conflict—be it between academic disciplines or different political worldviews.

2 From Information Society to 'Big Data'

Since the advent of the term 'Big Data,' society has become increasingly aware of its ubiquitous informatization. However, an initial discourse about the risks and potentials of mass data, its analytics and its social implications emerged as early as in the 1960s—from Marshal McLuhan's [1] vision of an electronically networked "global village" and Tadao Umesao's [2] idea of an upcoming "information society" to Arthur Miller's [3] concern over an "assault on privacy", Alvin Toffler's ([4]: 350) thesis of an "information overload", and Hal Beckers [5] widely noticed question "Can users really absorb data at today's rates? Tomorrow's?". In this respect, the recent debate can be seen as yet another step in a gradual process which started decades ago (cf. Table 1).

Today, the umbrella term 'Big Data' covers a wide range of data mining methods and application areas: insurances

Special Issue "Big Data", edited by Ullrich Meyer and Kristian Kersting.

✉ Johannes Weyer
johannes.weyer@tu-dortmund.de

[1] FernUniversität in Hagen, Hagen, Germany

[2] University of Stuttgart, Stuttgart, Germany

[3] WZB Berlin Social Science Center, Berlin, Germany

[4] TU Dortmund University, Dortmund, Germany

[1] http://www.abida.de.

Published online: 20 December 2017

Table 1 Phases in the informatization of society

1960s/1970s	1980s/1990s	2000s	2010+
Emergence of the concept and term 'information society'	Beginning informatization of everyday life (e.g., videotex, WWW)	Rise of data-focused technology companies and society-wide 'Web 2.0' discourse	'Big Data': Growing societal awareness of informatization

are applying individualized pricing models based on self-tracking, promising equitability but bearing the risk of an erosion of solidarity. Predictive policing programs are set to identify potential criminal activity and, at best, to prevent crime. Military intelligence aims to integrate machine data generated by ships, vehicles, aircraft, drones or satellites, human generated data from social media sites and business data from e-commerce transactions. Political data analysis has led to sophisticated micro-targeting and predictive analytics. Mass data analysis is seen to profoundly change healthcare systems and the treatment of disease. Data-orientated industrial management is trusted to increase productivity, quality as well as agility. And 'smart cities' are promoted to integrate Big Data and the Internet of Things to improve the efficiency of urban services [6].

3 Big Data Analytics and Its Societal Implications

Data evaluation strategies vary by application and implicate heterogeneous societal challenges. Quite a few established modes of Big Data analytics—e.g., smart traffic management—operate at higher levels of aggregation and do not necessarily resort to personally identifiable information. In contrast, various practices of behavior tracking and profiling challenge traditional notions of privacy as they aggregate and combine detailed individual data, which eventually results in accurate predictions of "highly sensitive personal attributes including: sexual orientation, ethnicity, religious and political views, personality traits, intelligence, happiness, use of addictive substances, parental separation, age, and gender" ([7]: 5802). Thus, the current hegemony of a few multinational companies as the operators of key digital infrastructures poses a challenge to state control and regulation.

For one, ubiquitously used services such as Twitter or Facebook contribute—with their predefined filter algorithms—significantly to the structuring of communication and social life. Secondly, digital infrastructures are opening up expanded possibilities of observation and control as the profiles of their users can be evaluated and sanctioned much more efficiently than before. This applies in particular to the ecosystems of smartphones, tablets and wearables: With the 'walled gardens' of Apple's iOS and Google's Android devices, the production and use of content or software have indeed become simpler. However, this standardization has also served to buttress the rule-setting force of a few dominant operators which is accompanied by an unprecedented control over user data [8].

Table 2 Distinct research cultures (Source: [9], modified)

	Social science	Engineering/data science
Goals	Search for explanation *why* something happens; develop theory	Search for accurate prediction of *what* happens; create algorithm
Focus	Explanation	Practical relevance
Beliefs	Data are biased	Data have ground truth cases

Against this backdrop, social science has an important role to play in the Big Data discourse: On a general level, sociologists want to generate *generalizable knowledge* and explanations, while engineers and data scientists tend to aim for structured and *applicable information* (cf. Table 2). Engineers and data scientists often focus on creating products that work at solving an applied problem and usually believe "in a 'ground truth' to which they can train models and align their solutions" ([9]: 25); in contrast, social scientists aim to contextualize present societal dynamics both within broader socio-economic developments as well as long-term social transformation processes in order to explain why something happens.

In this sense, social science is set to act as an imperative counterbalance to popular persuasions that "with enough data, the numbers speak for themselves" [10]—in particular as cheap and large data sets bear the risk to solely rely on quantifiable data and even the most detailed data sets remain objects of subjective or context-dependent interpretation. The bigger the data sets, the greater is though the probability to discover randomly congruent patterns and to deviate spurious correlations.[2]

In order to highlight the societal implications of contemporary practices of data collection and use, we now turn to individual data-related practices, before shedding light on the governance of complex systems.

[2] See, e.g., http://tylervigen.com/spurious-correlations.

4 Self-Tracking of Private Life

By means of smartphones and wearables (smart watches, activity trackers, smart glasses, etc.), individual behavior and bodily functions are now accessible for informatization, too. Different sensors, which are mostly worn on the body, allow their users to collect continuous data sets in an automated manner. In the past, this was reduced to selective data points or required a significant effort ([11]: 9–10).

From a sociological perspective, self-tracking is instructive as it allows to study the impact of informatization in an empirical field, that has not been determined and regulated by numbers to the same extent in the past. Additionally, the users actively form part of the production of sensible data sets. Some aspects of this specific data generation process shall be introduced subsequently, focusing on the enhancement of individual knowledge, comparison, and gamification.

4.1 Enhancement of Knowledge

"Knowledge through numbers" is the slogan of the Quantified Self-center (QS),[3] which started in 2007 and has become a world-wide movement which unites users, developers and other people interested in the informatization of personalized information. The enhancement of individual knowledge and self-optimization [12] are key factors for self-tracking, regardless of the initial motivation (health, sporting performances, lifestyle). Interestingly, these goals are frequently reduced to rather 'mundane' objectives, such as 'looking better', 'being a bit healthier', or 'loosing five kilos' [13]. General motivations for self-tracking can be classified as following [14]:

- *Self-reflection* As a consequence of widespread uncertainty of modern life ([15]: 199) self-tracking has become a new source of meaning and social recognition.
- *Optimization* Self-assessment contributes to a better understanding of the own body and helps to identify options for optimization ([12]), e.g. with the aims of curing diseases, controlling emotions, anticipating risks or increasing performance. Instant feedback of smart devices allows for reflexive control and a 'flexible' management of behavior (e.g., 'flexible dieting' or 'if it fits your macros').
- *Emancipation* Self-trackers see themselves as inventive and competent patients, who collect and compare health data and monitor their bodies independently. This autonomous access to body knowledge reduces their dependency on orthodox medicine and allows them to escape the informational hegemony of medical experts ([8]: 3).
- *Compliance* New social norms, which shift the responsibility for one's own body to the individual, have additionally promoted the practices of self-assessment.

4.2 Comparison and Gamification

Self-trackers mostly see themselves as first point of reference; the only significant 'sample', which does not require any comparisons or average values, is the own body [16].

In contrast to this very individualized data-practice, various applications also offer possibilities to share one's data with others and to perform permanent comparisons, opening up the possibility to gain badges and prizes or to access additional levels. Gamification approaches, which apply game-typical elements on tasks of everyday life and work, are mainly used for motivational aims ([11]: 23f). These extrinsic incentives lead to the (voluntary) discloser of data to the respective platform providers or create a certain group pressure to participate in self-tracking or sharing one's data ([11]: 115ff). Furthermore, self-tracking thus can be seen as the attempt to create a collective understanding of the categorization, evaluation, and comparison of the modern self [17].

Consequently, self-tracking provides new types of data insights into previously not datafied fields of bodily and everyday performances and enables new life styles including direct behavioral control. However, these practices can be regarded critically regarding data privacy, surveillance and societal power relations.

5 Real-Time Control of Complex Systems

Digital mass data not only allow to monitor individuals and to control individual behavior. Furthermore, companies and state authorities that manage large-scale socio-technical systems (e.g., transport or energy systems), are increasingly able to control those systems in real-time, if they have access to real-time data from the respective sensors, machines, users or customers [18].

While driving modern cars or using public transport, every user who has agreed to the terms of service transmits large amounts of data concerning position, speed, destination, etc. These data sets are used by traffic control centers to compute an up-to-date picture of the whole system and to identify patterns such as traffic jams or congestions. Yet, the main feature is prediction: Route guidance systems (e.g., Google Maps, TomTom) utilize the recorded patterns to predict future states of the system and to recommend bypasses or other strategic options. Users are free to comply with

[3] http://quantifiedself.com/.

the advice or to make own choices. Hence, the state of the overall transportation system a few minutes later emerges as the aggregated result of a huge number of decentralized decisions of independent actors. However, they are subject to control of a central authority that manages the system using soft incentives. This central authority pursues global goals, e.g. avoiding traffic jams or reducing CO_2 emissions. This may contradict individual objectives of moving fast, in a comfortable and self-determined way.

In the cases of smart transportation and smart grids, we can thus observe a new mode of governance: a 'soft' or 'smart' central control of decentralized, distributed systems [19, 20]. Individual actors maintain their autonomy to search for local optimum while at the same time being part of a rather global process, automatically guided and optimized by algorithms in real-time. In this case, privacy concerns are less relevant, because most systems are based on anonymized data. It is rather the opacity of real-time algorithmic decisions that may create a conflict between different participants [21]. In addition it remains an open question, to what extent individuals, that have gotten used to route guidance or other algorithmic systems, are still capable of taking their very own decisions, if the algorithmic recommendation is obviously erroneous and significant time pressures emerge. Whereas e.g. road trips formerly required planning as preceded action (looking at the road map *before* the trip starts), real-time systems of today enable real-time planning (start your trip and follow the continuously updated route guidance system). Traffic users that get accustomed to these services thus may increasingly become dependent on them and lose options.

6 Political Regulation of Big Data

As previously shown, Big Data may entail undesired societal consequences. Critics state that, without proper regulation, Big Data threatens individual liberties and democratic principles.

Yet, a common claim is that Big Data and regulation are antithetical. Some authors argue that Big Data renders the differentiation between personally identifiable data and other data obsolete, thus challenging current regulation [22]. In a similar vein, critical scholars (and civil rights groups) sustain that many Big Data related practices undermine the data protection principles of data minimization and purpose specification or contextual integrity [23]. A strict implementation of the current data protection regulation would finally stifle Big Data based innovation, as many industrials, politicians and researchers argue.

The perceived tension between Big Data and current regulation has led to a multitude of different regulatory suggestions [24].

- In the domain of *state regulation*, the most important ideas are to adapt data protection (e.g., by introducing vast data portability rights), and to improve the implementation of current data protection, for example via privacy impact assessments or better resources for independent data protection authorities. Another set of propositions calls for the complementation of data protection with other regulatory approaches, such as anti-trust regulation, and to find ways to protect human, civil and consumer rights, for example through anti-discrimination policy.
- In the domain of *corporate self-regulation*, the main propositions concern corporate social responsibility (CSR) instruments such as codes, standards and norms. Another regulatory approach is to encourage the development of privacy-enhancing and fairness-enhancing technologies: privacy by default, privacy by design, equal opportunity by design, bias mitigation and others. In addition, companies can participate in Big Data regulation by certifying their products, practices and organizations via seals and audits.
- In the domain of *regulation through civil society*, class actions, for instance lawsuits filed by consumer protection agencies, and support for citizens and consumers have been recommended to cope with Big Data induced risks.
- In the domain of *professional self-regulation*, relevant professions, such as data science professionals and data protection officers, could autonomously develop ethical guidelines for practices and qualification.
- Finally, the *regulation through self-protection* encompasses recommendations that aim at informing and sensitizing citizens and users in order to activate individual and collective self-protection.

Thus, this multitude of regulatory proposals—which can overlap, be combined and be implemented in very different ways—reveals that the debate about the political regulation of Big Data is not a mere search for technical adaptations to technological innovations.

There is no quick regulatory answer to technological change. In some cases, technological answers might be available and helpful. Yet in many cases, the conflict between different propositions is a struggle between different ideas and actors with various interests, resources and worldviews [25]. For instance, a neo-liberal conception, pleading for corporate self-regulation through ethical codes and improved measures for informed consent by users [26] is diametrically opposed to a (socio-)democratic view that calls for stronger state intervention in Big Data regulation through bans [27], the provision of public internet and data services [28], and an approach to data protection that is centered on collectives and not merely on individuals [29]. However, all

those propositions share the challenge to establish an effective regulatory regime on a transnational level—a domain where states have limited regulatory capacity.

7 Conclusion

The informatization of almost every aspect of life has profound societal implications. Smart devices as smartphones and wearables, smart cars or smart meters collect and transmit a previously unknown stream of digital mass data. Consumers have become an active part of the game by voluntarily supplying personal content, private usage information and various kinds of hidden meta data to the respective providers, in turn gaining instant analytics (e.g. self-tracking), real-time recommendations and knowledge-representation (e.g. route planning) or constant feedback on common tasks and processes.

From a sociological point of view, this increasing informatization of everyday life and the digitalization of societal and economic sectors first and foremost raises—aside from privacy concerns—the issue of control and regulation: since aggregated and interconnected data can be readily used to influence or even control individual and collective behavior as well as complex socio-technical systems, the currently observable infrastructural power of a few globally leading companies such as Apple or Google, controlling encompassing platforms of data collection and distribution [30], strongly calls for regulation. A future regulation should combine the implementation of suitable legal instruments on a transnational level with a constant monitoring of the societal effects of Big Data. Only when we know under what circumstances Big Data leads to discrimination, manipulation and other undesired effects, will we be able to criticize and to prevent them. For this aim, social science, data science and engineers should join forces.

References

1. McLuhan M (1962) The Gutenberg galaxy. University of Toronto Press, Toronto
2. Umesao T (1963) Information industry theory: dawn of the coming era of the ectodermal industry. Hoso Asahi 1/1961:4–17
3. Miller AR (1967) The assault on privacy. University of Michigan Press, Ann Arbor
4. Toffler A (1970) Future shock. Bantam Books, New York
5. Becker HB (1986) Can users really absorb data at today's rates? Tomorrow's? Data Commun 15(8):177–193
6. Seyfert R, Roberge J (eds) (2017) Algorithmic cultures: essays on meaning, performance and new technologies. Routledge, London/New York
7. Kosinski M, Stillwell D, Graepel T (2013) Private traits and attributes are predictable from digital records of human behavior. Proc Natl Acad Sci 110(15):5802–5805
8. Dolata D, Schrape JF (2016) Masses, crowds, communities, movements: collective action in the internet age. Soc Mov Stud 15(1):1–18
9. McFarland DA, Lewis K, Goldberg A (2016) Sociology in the era of Big Data: the ascent of forensic social science. Am Sociol 47(1):12–35
10. Anderson C (2008) The end of theory: the data deluge makes the scientific method obsolete. Wired 16(7). https://www.wired.com/2008/06/pb-theory/. Accessed 10 Dec 2017
11. Lupton D (2016) The quantified self. Polity Press, Cambridge and Malden
12. Ruckenstein M, Pantzar M (2017) Beyond the quantified self: thematic exploration of a dataistic paradigm. New Media Society 19(3):401–418
13. Didžiokaitė G, Saukko P, Greiffenhagen C (2017) The mundane experience of everyday calorie trackers: beyond the metaphor of quantified self. New Media Soc. https://doi.org/10.1177/1461444817698478
14. Selke S (2016) Rationale Diskriminierung durch Lifelogging—Die Optimierung des Individuums auf Kosten des Solidargefüges? In: Andelfinger VP, Hänisch T (eds) Die Optimierung des Individuums—wie weit wollen, sollen, dürfen wir gehen? Springer, Berlin, pp 53–71
15. Selke S (2014) Lifelogging: Wie die digitale Selbstvermessung unsere Gesellschaft verändert. Ullstein eBooks
16. Gerlek S, Kappler KE, Noji E (2017) Digitale Medien in leiblichen Praktiken. In Rebane G (ed) Identität und kulturelle Praktiken im digitalen Zeitalter. Könighausen & Neumann, Würzburg
17. Vormbusch U (2016) Taxonomien des Selbst. Zur Hervorbringung subjektbezogener Bewertungsordnungen im Kontext ökonomischer und kultureller Unsicherheit. In: Duttweiler S, Gugutzer R, Passoth JH, Strübing J (eds) Leben nach Zahlen. Self-Tracking als Optimierungsprojekt? Transcript, Bielefeld, pp 45–62
18. Weyer J (2014) Einleitung: Netzwerke in der mobilen Echtzeitgesellschaft. In: Weyer J (ed) Soziale Netzwerke. Konzepte und Methoden der sozialwissenschaftlichen Netzwerkforschung, 3rd edn. Oldenbourg, München, pp 3–37
19. Rochlin GI (1997) Trapped in the net. The unanticipated consequences of computerization. Princeton, Princeton UP
20. Lessig L (2006) Code: version 2.0. Bacis Books, New York
21. Saurwein F, Just N, Latzer M (2015) Governance of algorithms: options and limitations. Info 17(6):35–49. https://doi.org/10.1108/info-05-2015-0025
22. Ohm P (2014) Changing the rules: general principles for data use and analysis. In: Lane J, Stodden V, Bender S, Nissenbaum H (eds) Privacy, big data, and the public good. Cambridge University Press, New York, pp 96–111
23. Barocas S, Nissenbaum H (2014) Big Data's end run around anonymity and consent. In: Lane J, Stodden V, Bender S, Nissenbaum H (eds) Privacy, Big Data, and the public good. Cambridge University Press, Cambridge, pp 44–75
24. Ulbricht L, Haunss S, Hofmann J, Klinger U, Passoth JH, Pentzold C, Schneider I, Strassheim H, Voß JP (2018) Dimensionen von Big Data: eine politikwissenschaftliche Systematisierung". In: Heil R, KolanyRaiser B, Orwat C (eds) Big Data und Gesellschaft. Eine multidisziplinäre Annäherung. Springer, Wiesbaden **(forthcoming)**
25. Ulbricht L (2017) Machtkämpfe um Big Data: Bürger und Konsumenten müssen geschützt werden. WZB-Mitteilungen 155:18–21
26. Foster JB, McChesney RW (2011) The internet's unholy marriage to capitalism. Mon Rev 62(10):1–30
27. Roßnagel A, Geminn C, Jandt S, Richter P (2016) Datenschutzrecht 2016 "Smart" genug für die Zukunft?: Ubiquitous computing und Big Data als Herausforderungen des

Datenschutzrechts. http://www.uni-kassel.de/upress/online/OpenAccess/978-3-7376-0154-2.OpenAccess.pdf. Accessed 10 Dec 2017

28. Lewandowski D (2014) Why we need an independent index of the web. In: König R, Rasch M (eds) Inc Reader: vol 9. Society of the query reader. Reflections on web search. Inst. of Network Cultures, Amsterdam, pp 49–58
29. Hull G (2015) Successful failure: what Foucault can teach us about privacy self-management in a world of Facebook and big data. Ethics Inf Technol 17(2):89–101. https://doi.org/10.1007/s10676-015-9363-z
30. Dijck J (2016) #AoIR2016: opening keynote "The Platform Society" by José van Dijck. https://www.youtube.com/watch?v=-ypiiSQTNqo. Accessed 10 Dec 2017